AD-A039 195 NORTH CAROLINA UNIV AT CHAPEL HILL F/G 12/1
ESTIMATING THE MEAN OF A CORRELATED BINARY SEQUENCE WITH AN APP--ETC(U)
APR 77 G S FISHMAN, L R MOORE N00014-76-C-0302

UNCLASSIFIED TR-77-2 NL

| OF |
ADA039 195

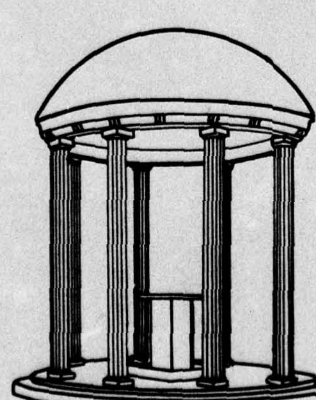
END

DATE
FILMED
6 - 77

(12)

OPERATIONS RESEARCH AND SYSTEMS ANALYSIS

UNIVERSITY OF NORTH CAROLINA
AT CHAPEL HILL



ESTIMATING THE MEAN OF A CORRELATED BINARY SEQUENCE WITH AN APPLICATION TO DISCRETE EVENT SIMULATION

George S. Fishman and Louis R. Moore

Technical Report 77-2
April 1977

Curriculum on Operations Research
and Systems Analysis

University of North Carolina at Chapel Hill

This research was supported by the Office of Naval Research under contract N00014-76-C-0302. Reproduction in whole or in part is permitted for any purpose of the United States Government.

See 1473

Abstract

This paper discusses a procedure for interval estimation of the mean Θ of a correlated binary (0,1) sequence. The method assumes that the sequence is strictly stationary and that a particular string of m binary digits is a recurrent event in the sequence, where $m \geq 1$ is unknown. Of the 2^m choices for the possible recurrent events, the strings of all zeros and of all ones are examined.

For each m=1,2,... the sequence is demarcated by entrance to the recurrent event. The subsequences between the demarcation points thus form independent epochs by assumption. Classical techniques then yield variance estimates for the number of ones and zeros in the epoch as well as an estimate of the covariance of the ones and zeros. A quadratic equation in Θ is solved to obtain an interval estimate.

Each string of all ones or all zeros examined yields a $1-\alpha$ confidence interval. The intervals are intersected to obtain shorter intervals with confidence greater than $1-2\alpha$. Since each m=1,2,... yields an interval, a conservative rule is developed to determine the m whose interval is finally used. This rule is based upon the empirical run lengths in the binary sequence.

The procedure is then applied to interval estimation of the fractile for the waiting time distribution in a simulation of the M/M/1 queue with activity level 0.9. For Θ=0.1 and 0.5 the proposed method worked well. For Θ=0.9 results showed some degradation. An error analysis led to a set of recommendations for keeping performance in practice close to the desired theoretical levels. An appendix describes algorithms for computing the critical quantities upon which the proposed method relies.

1. Introduction

Although many techniques exist for the statistical analysis of simulation output in general [11], little has appeared in the literature that specifically addresses the estimation of fractiles of a distribution that arises in a simulation. This problem is not unique to simulation. In its more general form it concerns the estimation of the mean of a random binary (0,1) sequence whose elements are correlated. The purpose of this paper is to describe a method of interval estimation for the more general problem and illustrate the method with a simulation example.

The proposed method relies on the theory of recurrent events, as described in Feller [9]. Although application of these methods to simulation output analysis is not new, the approach taken here is. Crane and Iglehart [7 , 8] and Fishman [11] exploit the regenerative properties of certain simulation models to analyze output. The theory of regenerative processes generalizes the theory of recurrent events to include continuous as well as discrete time. See Smith [13] for details. What is new here is the application of the theory of recurrent events without reference to particular simulation models. Specifically, the proposed method enables one empirically to identify a state that appears to have the properties of a recurrent state and then to use that state to cut up the sample path into approximately independent identically distributed segments that can then be used with relatively elementary statistical methods to compute an interval estimate. Moreover, the binary nature of the data offers considerable computational conveniences that make this method of inference attractive.

Whenever one deals with discrete finite state data, the inclination to approximate the sequence by a Markov chain is hard to resist. Once the

order of the chain is determined, the subsequent inference is well known [1]. Moreover, every state of the given order is a recurrent state. To determine the order of the chain one can apply chi-square or likelihood ratio methods the theory for which is also known in principle [1]. Although our study began in this way, we quickly learned that the computational problems that arose in estimating the order were considerably greater in complexity than those of the less restrictive recurrent event approach. Moreover, the performance was not nearly as good as with the recurrent event approach. Therefore we present here the results for the recurrent approach, even though the statistical inference for this theory is not as complete in the statistical literature as for Markov chains.

Section 2 introduces the reader to the problem as it is formulated in the theory of recurrent events. Section 3 describes how one can obtain shorter interval estimates by using results based on analyses for different recurrent states. Section 4 indicates which among the many potential states deserve consideration. Section 5 describes estimators that need to be substituted for population parameters and Section 6 contains a computing scheme that facilitates efficient computation. Section 7 presents results that show how the method performs for the .1, .5, and .9 fractiles of the waiting time distribution in an M/M/1 queueing simulation with activity level of .9. Finally, Section 8 contains recommendations for interpretation and utilization of interval estimates that a user of the proposed method may encounter in practice.

2. The Recurrent Model

Let $\{X_i;\ i = 1,\ldots,n\}$ denote a sequence of observations on a strictly stationary binary (0,1) stochastic sequence. Define for $m < n$

(1) $$Y_i^{(m)} = \sum_{j=0}^{m-1} 2^j X_{i-j} \qquad i=m,\ldots,n.$$

Suppose there exists an m and $0 \le y^* \le 2^m - 1$ such that

(2) $$pr(X_k = x \mid X_j = z, Y_i^{(m)} = y^*) = pr(X_k = x \mid Y_i^{(m)} = y^*)$$

$$j \le i \le k; \qquad n \ge i \ge m; \qquad x,z = 0,1.$$

One can then apply the theory of recurrent processes [13] to the analysis of $\{X_i\}$, provided $Y_i^{(m)} = y^*$ infinitely often as $n \to \infty$. In particular, let

$$N_i^{(m)} \equiv \sum_{j=m}^{i} \delta(Y_j^{(m)} - y^*) \qquad i=m,\ldots,n$$

$$\delta(x) \equiv \begin{cases} 1 & x=0 \\ 0 & x \ne 0 \end{cases}$$

(3) $$T_j^{(m)} \equiv \min\{i \ge m: N_i^{(m)} = j\} \qquad j=1,\ldots,N_n^{(m)}$$

$$T_0^{(m)} \equiv 1, \qquad T_{N_n^{(m)}}^{(m)} \equiv n+1.$$

Here $\{T_j^{(m)}\,;\ j=1,\ldots,N_n^{(m)}\}$ defines a discrete renewal process where $T_j^{(m)}$ is the time of the jth renewal.

Let

$$C_j^{(m)} = T_{j+1}^{(m)} - T_j^{(m)}$$

(4) $$S_j^{(m)} = \sum_{i=T_j^{(m)}}^{T_{j+1}^{(m)}-1} X_i \qquad j=0,\ldots,N_n^{(m)}.$$

Then $\{C_j^{(m)}\,;\ j=1,\ldots,N_n^{(m)}-1\}$ and $\{S_j^{(m)}\,;\ j=1,\ldots,N_n^{(m)}-1\}$ are each sequences of i.i.d. non negative integer valued random variables. Let

$E(C_j^{(m)}) = \mu_C^{(m)}$, $E(S_j^{(m)}) = \mu_S^{(m)}$, $\mathrm{var}(C_j^{(m)}) = \sigma_{CC}^{(m)}$, $\mathrm{var}(S_j^{(m)}) = \sigma_{SS}^{(m)}$

for $j = 1,\ldots,N_n^{(m)} - 1$,

(5)
$$\mathrm{cov}(C_j^{(m)}, S_k^{(m)}) = \begin{cases} \sigma_{CS}^{(m)} & j=k=1,\ldots,N_n^{(m)}-1 \\ 0 & j \neq k \end{cases}$$

and $\Theta \equiv E(X_i)$. Then

(6)
$$\hat{\Theta} = n^{-1} \sum_{j=1}^{n} X_i$$

is an unbiased estimator of Θ. Since one can show that $\Theta = \mu_S^{(m)}/\mu_C^{(m)}$, it is instructive to study the linearized observations

(7)
$$Z_j^{(m)} \equiv S_j^{(m)} - \Theta C_j^{(m)} \qquad j=0,\ldots,N_n^{(m)}$$

and the sum

(8)
$$Z^{(m)} \equiv \sum_{j=1}^{N_n^{(m)}-1} Z_j^{(m)}$$

for which

(9a) $E(Z^{(m)}) = E(N_n^{(m)} -1)E(S_j^{(m)}) - \Theta E(N_n^{(m)} -1)E(C_j^{(m)}) = 0$

(9b) $\mathrm{var}(Z^{(m)}) = E(N_n^{(m)} -1)(\sigma_{SS}^{(m)} - 2\Theta\sigma_{CS}^{(m)} + \Theta^2\sigma_{CC}^{(m)})$.

Also the distribution of $Z^{(m)} / \sqrt{\mathrm{var}(Z^{(m)})}$ approaches $N(0,1)$. This limiting distribution provides a method for constructing an interval estimate for Θ for large n.

For expository convenience let us momentarily suppress the superscript (m). Let

(10)
$$f(\Theta) = [C^2 - Q^2E(N_n-1)\sigma_{CC}]\Theta^2 - 2[CS - Q^2E(N_n-1)\sigma_{CS}]\Theta + [S^2 - Q^2E(N_n-1)\sigma_{SS}],$$

$$(11) \qquad C \equiv \sum_{j=1}^{N_n - 1} C_j \qquad\qquad S \equiv \sum_{j=1}^{N_n - 1} S_j \ ,$$

Q being the $1-\alpha/2$ value of the unit normal distribution. Then for large n, $pr[f(\Theta) \leq 0] \doteq 1-\alpha$. Solutions to the probability argument yield the seven cases in Table 1. Here a_1 and a_2 are the roots of $f(\Theta)$

Table 1

$1-\alpha$ Interval Estimate of Θ

Case	Interval	f(0)	f'(0)	f(1)	f'(1)
1	$[a_1,a_2]$	+	-	+	+
2	$[a_1,1]$	+	-	-	$\pm$
3	$[0,a_2]$	-	$\pm$	+	+
4	$[0,a_1]\cup[a_2,1]$	-	+	-	
5	$[0,a_1]$	-	+	+	$\pm$
6	$[a_2,1]$	+	$\pm$	-	-
7	$[0,1]$		$f(\Theta) \neq 0$	$\forall\Theta\epsilon[0,1]$	

with $|a_1| \leq |a_2|$. One can easily show that as $n \to \infty$ case 1 becomes the type of interval estimate obtained. This procedure for computing interval estimates is due to an extension of the work in Bliss [2,3] by Fieller[10].

3. Many Recurrent States

In practice many stochastic processes with a discrete state space have more than one recurrent state. In particular, if the sequence of states $\{y_j = j-1;\ j=1,\ldots,2^m\}$ were all recurrent then $\{X_i\}$ would be an mth order Markov chain. Although an analyst is free to choose any of the set of recurrent states for the computation of an interval estimate each state produces an interval of different width. Moreover, the n for

which the aforementioned asymptotic properties usually hold differs with each selected state.

Suppose that for a given $\{X_j\}$ one computes a set of k interval estimates for θ. For expository convenience assume that all the intervals correspond to case 1 in Table 1 and that $[a_1(i),a_2(i)]$ is the ith $1-\alpha$ interval estimate of θ for $i=1,\ldots,k$. Then, using Bonferroni's inequality [12] one can show that

$$(12) \qquad \prod_{i=1}^{k} pr[a_1(i) \leq \theta \leq a_2(i)] \geq 1-k\alpha \, .$$

If these intervals intersect then the intersection provides an interval estimate for θ with probability exceeding $1-k\alpha$.

Consider the case $k=2$ for which $a_1(1)\leq a_1(2)\leq a_2(1)\leq a_2(2)$. Then $[a_1(2),a_2(1)]$ includes θ with probability of at least $1-2\alpha$. Here a question arises as to whether a shorter interval would be obtained by computing a $1-2\alpha$ interval estimate for $i=1$ or $i=2$ separately. Section 8 examines this issue with regard to a specific example.

4. States to Consider

As the introduction indicates, our intention is to describe a procedure for computing interval estimates for θ by approximating $\{X_j\}$ by a recurrent process. Since the state vector as defined in (1) implies 2^m potential recurrent states, even a moderate m yields an excessive number of options. However, two particular states deserve special attention. The state $y^* = 0$ implies that when $Y_i=y^*$, $X_{i-j}=0$ for $j=1,\ldots,m$. Now it is plausible that for sufficiently large m, $Y_i=y^*$ contains all conditioning information about the past and therefore this y^* of order m enables one to

demarcate $\{X_j\}$ into sequences of i.i.d random variables as in (4). Alternatively the state $y^* = 2^m - 1$ implies that when $Y_i = y^*$, $X_{i-j} = 1$ for $j=1,\ldots,m$. A similar plausibility argument applies here.

In the remainder of this paper we restrict attention to these potential regenerative states. Doing so relieves us of the burden of considering the remaining 2^m-3 states for each value of m considered but leaves us open to the possibility of excessive computation if m turns out to be large. Happily the binary character of the data makes this possibility remote as the computing scheme in Section 6 shows.

5. Sample Variances

In practice one does not know $E(N)$, σ_{CC}, σ_{SS} and σ_{CS}. To resolve this problem we consider the sample quantities

$$s^2(C) = \sum_{j=1}^{N_n-1} (C_j - \bar{C})^2$$

$$s^2(S) = \sum_{j=1}^{N_n-1} (S_j - \bar{S})^2$$

$$(13) \qquad s(C,S) = \sum_{j=1}^{N_n-1} (C_j - \bar{C})(S_j - \bar{S})$$

$$\bar{C} = \frac{1}{N_n-1} \sum_{j=1}^{N_n-1} C_j \, , \qquad \bar{S} = \frac{1}{N_n-1} \sum_{j=1}^{N_n-1} S_j .$$

Since

$$s^2(C) = \sum_{j=1}^{N_n-1} (C_j - \mu_C)^2 - \frac{\left[\sum_{j=1}^{N_n-1} (C_j - \mu_C)\right]^2}{N_n-1}$$

One can use the well known results of renewal theory [5,13]

$$(14a) \qquad E(N_n) \sim n / \mu_C$$

(14b) $$\mathrm{var}(N_n) \sim n\sigma_{CC} / \mu_C^3$$

in a series expansion to show that

(15) $$E[s^2(C)] = E(N_n-1)\sigma_{CC} - 2\sigma_{CC} + O(1/n).$$

Similarly

(16)
$$E[s^2(S)] = E(N_n-1)\sigma_{SS} - \sigma_{SS} - \Theta^2\sigma_{CC} + O(1/n)$$
$$E[s(C,S)] = E(N_n-1)\sigma_{CS} - \sigma_{CS} - \Theta\sigma_{CC} + O(1/n).$$

These results reveal that $s^2(C)$, $s^2(S)$ and $s(C,S)$ are asymptotically unbiased estimations of $E(N_n-1)\sigma_{CC}$, $E(N_n-1)\sigma_{SS}$ and $E(N_n-1)\sigma_{CS}$ respectively. Moreover, one can substitute these quantities into (9b) and show that the resulting distribution of $Z^{(m)}/\sqrt{\mathrm{var}\, Z^{(m)}}$ again approaches the standard normal law as $m \to \infty$.

6. A Computing Scheme†

For analysis, binary data offer conveniences that allow considerable computational efficiency. The sequence $\{X_i\}$ consists of runs of ones and zeros of varying lengths

(17)
$$L_{2j-1} = \text{length of the jth run of zeros}$$
$$L_{2j} = \text{Length of the jth run of ones}$$

where $L_1 \equiv 0$ if $X_1=1$. If M runs occur then $\{L_1,\ldots,L_M\}$ summarizes $\{X_i\}$ without any information loss. One can easily establish a bound on $E(M)$. Suppos $\{X_i\}$ i a sequence of i.i.d random variables with $pr(X_i=1)=p$. Then $E(M) = np(1-p) \le n/4$. If $p=.9$ $E(M) = 0.09n$.

† The Appendix contains algorithms for computing the critical quantities described here.

Now in the cases at hand we anticipate positive correlation between elements of $\{X_i\}$. Therefore one expects runs to be longer and be smaller for a given n.

The computing scheme presented here uses the recurrent states $y^* = 2^{km}-1$ for $k = 0$ and 1. In the remainder of this section we assume k and m given and suppress reference to them except where clarity calls for explicit mention. Let

$$K_0 \equiv k - 1$$

$$(18) \qquad K_j \equiv \min\,(2i + k-1 > K_{j-1} : L_{2i+k-1} \geq m) \qquad j = 1,2,\ldots\,.$$

For computational convenience we set $k_j \equiv M+1$ if the aforementioned minimum does not exist. Define

$$(19) \qquad J \equiv \min\,(j : K_j = M+1) - 1$$

Then $\{K_j;\ j=1,\ldots,J\}$ forms a sequence of <u>cutpoints</u> for given m and k. As m increases for given k the number of cutpoints decreases, which accelerates subsequent search procedures.

For given k and m let

$$(20) \qquad B_J = \sum_{i=1}^{J} L_{K_i} - Jm$$

$$
\left.\begin{aligned}
B_j &= \sum_{i=1}^{j} L_{K_i} - jm \\
D_j &= \sum_{K_j<i<K_{j+1}} L_i + m \\
E_j &= \sum_{K_j<2i+k-1<K_{j+1}} L_{2i+k-1} + m
\end{aligned}\right\} \quad j=1,\dots,J-1 . \tag{21}
$$

Then one can show

$$
\begin{aligned}
C &= B_J + \sum_{j=1}^{J-1} D_j \\
S &= kB_J + (1-k)\sum_{j=1}^{J-1} D_j + (2k-1)\sum_{j=1}^{J-1} E_j \\
N_n &= B_J + J \quad ,
\end{aligned} \tag{22}
$$

where (11) and (3) define S, C and N_n. One can also show

$$
\begin{aligned}
s^2(C) &= B_J + \sum_{j=1}^{J-1} D_j^2 - C^2/(N_n-1) \\
s^2(S) &= kB_J + \sum_{j=1}^{J-1}\{kE_j^2 + (1-k)(D_j-E_j)^2\} \\
&\qquad - \{kS^2 + (1-k)(C-S)^2\}/(N_n-1) \\
s(C,S) &= kB_J + \sum_{j=1}^{J-1} D_j\{kE_j + (1-k)(D_j-E_j)\} \\
&\qquad - C\{kS + (1-k)(C-S)\}/(N_n-1) \ ,
\end{aligned} \tag{23}
$$

where (13) defines $s^2(C)$, $s^2(S)$ and $s(C,S)$.

The formulae for computing these quantities apply for $k = 0$ and 1. Also notice that the number of calculations J is always less than n and usually considerably smaller. An additional convenience arises for $m < L_{2j+k-1}$ for $j=1,\dots,J^{(m,k)}$. Then for scheme m+1 and k one has

$$
(24)\qquad
\begin{aligned}
J^{(m+1,k)} &= J^{(m,k)} \\
K_j^{(m+1,k)} &= K_j^{(m,k)} \\
B_j^{(m+1,k)} &= B_j^{(m,k)} - j \\
D_j^{(m+1,k)} &= D_j^{(m,k)} + 1 \\
E_j^{(m+1,k)} &= E_j^{(m,k)} + 1 \\
N_n^{(m+1,k)} &= N_n^{(m,k)} - J^{(m,k)} .
\end{aligned}
$$

7. Estimating m

The foregoing scheme for estimating $E(N_n-1)\sigma_{CC}$, $E(N_n-1)\sigma_{SS}$ and $E(N_n-1)\sigma_{CS}$ leaves one remaining, but critical, problem before one can compute an interval estimate for Θ. This concerns the estimation of m for given k. Firstly, m for $k=0$ usually differs from m for $k=1$. Secondly, a criterion is needed to determine a satisfactory m for each case. Consider the linearized form (7) with variance in (9b). Since

$$(25)\qquad s^2(Z^{(m,k)}) = s^2(S^{(m,k)}) - 2\hat{\Theta}s(C^{(m,k)},S^{(m,k)}) + \hat{\Theta}^2 s^2(C^{(m,k)})$$

provides an estimate of (9b) a conservative rule is†

$$(26)\qquad m^* = \min[m:\ s^2(Z^{(m,k)}) \geq s^2(Z^{(i,k)});\ i=0,\ldots,L_k^*]$$

$$L_k^* \equiv \sup_{1\leq i\leq \lceil M/2\rceil} (L_{2i+k-1}) .$$

Here L_k^* denotes the longest run of k's in the sample sequence $\{L_i\}$. This rule requires total enumeration of all possible run lengths to

† The quantity $\lceil x\rceil$ denotes the integer part of x.

determine m^*. Moreover, by selecting m^* based on the maximal $s^2(Z^{(i,k)})$ the rule picks the most conservative result that the empirical evidence will support. A less costly and less conservative rule is obtained as follows. Let m_i denote the ith order statistic of the sequence $\{L_{2i+k-1}, L_{2i+k-1} + 1: i=1,\ldots,\lceil M/2 \rceil\}$. Then the rule is defined as

$$(27) \qquad m^{**} = \min \left[m_j : s^2(Z^{(m_j,k)}) \geq s^2(Z^{(m_i,k)}),\ i=1,\ldots, 2\lceil M/2 \rceil \right].$$

This rule compares sample variances for the empirical runs of k's in the sample. These are a subset of the runs contained in (26).

Regardless of which rule is selected to estimate m, a possibility remains that the sample sequence $\{L_i\}$ does not contain runs of sufficient length to estimate m with accuracy. One way to check on adequacy is to plot $s^2(Z^{(i,k)})$ versus i and observe if the sample variance achieves an approximate plateau in the vicinity of its maximum. If it does then one can place confidence in the estimated m as providing suitable accuracy for the approximating scheme.

8. A Queueing Example

To illustrate the proposed estimation procedure we use a simulation of the M/M/1 queueing problem with arrival rate $\lambda = 1$ and service rate $\omega = 1/.9$.† This yields an activity level $\rho = .9$ which implies a high level of congestion. In particular the mean number of jobs in queue is $\rho/(1-\rho) = 9$ and the mean waiting time is $\rho/\lambda(1-\rho) = 9$. The objective is to estimate fractiles of the waiting time distribution F corresponding to waiting times $w = 0, 5.29, 19.78$. These fractiles are $\Theta = .1, .5, .9$, respectively, which cover a substantial part of the domain of F.

Let $\{W_i;\ i=1,\ldots,n\}$ denote a sample sequence of waiting times collected on a given simulation run, where collection begins after the effects of

† See Cox and Smith [6].

initial conditions has dissipated. Define

(28)
$$X_i = \begin{cases} 1 & W_i \le w \\ 0 & W_i > w \end{cases}$$

so that $\hat{\theta}$ in (6) estimates the fractile of F corresponding to w. Results in Blomquist [4] enable one to show that

(29)
$$\lim_{n\to\infty} n\,\mathrm{var}(\hat{\theta}) = F^*(w)\{1+\rho^2-2F^*(w)(1+\lambda(1-\rho)w)/(1-\rho)^2-1+F^*(w)\}$$

(30)
$$F^*(w) = \begin{cases} 1-\rho & w=0 \\ \rho e^{-w(\omega-\lambda)} & w>0 \end{cases}$$

so that one can compute the large sample variance of $\hat{\theta}$ for comparison with empirical results.

For each value of θ the sampling experiment consisted of $n=8192$ observations on 100 independent replication collected from a simulation that was operating in the steady state. Application of the procedures described in Sections 5,6 and 7 produced the results in Table 2. In particular, the theoretical mean is known <u>a priori</u>. The sample mean denotes the average values of $\hat{\theta}$ over the 100 replications. The theoretical variance follows from (29) divided by n whereas the sample variance is the average value of $s^2(Z^{(m^{**},k)})/(N_n^{(m^{**},k)}-1)n$ over the 100 replications. The coverage rates indicate the proportion of intervals based on Table 1 that include the theoretical θ.

Notice that the coverage rates for $\theta = .1$ and $.5$ are conservative whereas those for $\theta=.9$ need more careful scrutiny. Although the sample variances usually underestimate the corresponding theoretical variance, the width of the intervals appear larger than theoretically expected. The confusion disappears when we look at how the theoretical interval widths were computed. Asymptotically one can show that on a given run $\hat{\theta}$ has

Table 2

Results of Proposed Procedures for an

Approximating Recurrent State

$n=8192$, 100 replications, $1-\alpha=0.95$

	Mean	Variance x 10^{-4}	Coverage Rate	Interval Width	Order of Scheme m**	No. of Renewals $N_n - 1$
Theoretical	0.1000	2.809	0.95	0.0657	n.a.†	n.a.
Sample: k=0	0.1029	1.648	0.99	0.1269	33	4332
k=1	0.1029	1.937	0.97	0.1835	1††	991
Theoretical	0.5000	34.00	0.95	0.2286	n.a.	n.a.
Sample: k=0	0.5016	28.91	0.98	0.5050	25	2645
k=1	0.5016	28.48	0.95	0.4906	2	3915
Theoretical	0.9000	29.54	0.95	0.2131	n.a.	n.a.
Sample: k=0	0.9000	35.05	0.82	0.9299	13	718
k=1	0.9000	24.92	0.91	0.2257	38	6764

† Not applicable

†† Two replications chose m=0 and ninety-eight chose m=1. The choice m=1 is theoretically correct.

the normal distribution with mean θ and variance var($\hat{\theta}$), given in (29). Therefore $Q\sqrt{var(\hat{\theta})}$, which is the basis of the theoretical intervals in Table 2, provides centered intervals which have the shortest possible widths. Since the confidence intervals based on the sample data rely heavily on the procedure in Section 2, there is no reason to expect them to be the shortest possible. Nevertheless the mean width of .9299 for k=0 and θ =.9 indicates that, at least in this case, a scheme with k=0 has little focusing power. We examine this issue in greater detail shortly.

Notice that $m^{**} \doteq 1$ for θ =.1 and k=1. This agrees with theory since every time a job enters service immediately upon arrival, a renewal occurs. In particular the mean number of renewals for θ =.1 and w=0 is [5] $n(1-\rho)$ = 819.2, which does not differ substantially from the reported 991.

In section 3 we discussed a procedure for combining results for k=0 and k=1 to obtain shorter interval estimates. Table 3 lists the results based on (12) and compares them with the theoretically shortest achievable interval for $1-\alpha$ =.9, the smallest achievable theoretical probability. The dramatic reduction in widths compared to those in Table 2 is apparent. Notice

Table 3

Intersecting Confidence Intervals

θ		Coverage Rates	Interval Widths
0.1	Theoretical	.90	0.0551
	Sample	.96	0.0694
0.5	Theoretical	.90	0.1918
	Sample	.95	0.2472
0.9	Theoretical	.90	0.1788
	Sample	.74	0.1666

that the rates for θ = .1 and .5 remain. The poor performance for θ = .9 is

to be expected since Table 2 dictates a maximal achievable rate of .82.

We next discuss the poorer than expected performance for Θ =.9. Table 4 shows the frequency of intervals corresponding to the seven cases enumerated in Table 1. Notice that for Θ =.1 and .5 for k=0,1 the intervals occur principally among cases 1,2 and 3. For Θ =.9 and k=1 case 3 occurs exclusively. However for Θ =.9 and k=0 the less desirable cases 4 and 5 occur in 98 replications. In particular the 91 of case 5 offer insight into why the interval width for this case is so large in Table 2.

Table 4

Empirical Frequency of Interval Estimates by Case (n=8192)

Case	Interval	.1 k=0	.1 k=1	.5 k=0	.5 k=1	.9 k=0	.9 k=1
1	$[a_1,a_2]$	80	98	58	88	1	0
2	$[a_1,1]$	17	0	26	0	0	0
3	$[0,a_2]$	0	1	0	12	1	100
4	$[0,a_1]\cup[a_2,1]$	3	0	7	0	7	0
5	$[0,a_1]$	0	0	9	0	91	0
6	$[a_2,1]$	0	1	0	0	0	0
7	$[0,1]$	0	0	0	0	0	0

Here $f(\Theta)$ in (10) is inverted from the desirable situation that arises in cases 1,2 and 3.

Before passing final judgement it is instructive to investigate the effect of increased sample size on coverage rate and interval width for Θ=.9. Table 5 compares results for n=8192 and n=16384. Notice the substantial improvement in coverage rate for n=16384 and k=0. Regrettably no similar improvement occurs for the interval width. Moreover,

Table 5

Comparison of Results for θ=.9 and n=8192, 16384

	Mean	Variance x 10^{-4}	Coverage Rate	Interval Width	Order of Scheme m^{**}	No. of Renewals $N_n - 1$
n=8192						
Theoretical	0.9000	29.54	0.95	0.2131	n.a.	n.a.
Sample: k=0	0.9000	35.05	0.82	0.9299	13	718
k=1	0.9000	24.92	0.91	0.2257	38	6764
n=16384						
Theoretical	0.9000	14.77	0.95	0.1507	n.a.	n.a.
Sample: k=0	0.8996	12.00	0.95	0.9275	13	1362
k=1	0.8996	12.82	0.94	0.2028	38	13537
Intersecting Intervals						
n=8192						
Theoretical	0.9000	n.a.	≥.90	0.1788	n.a.	n.a.
Sample	0.9000	n.a.	0.74	0.1666	n.a.	n.a.
n=16384						
Theoretical	0.9000	n.a.	≥.90	0.1264	n.a.	n.a.
Sample	0.8996	n.a.	0.89	0.1518	n.a.	n.a.

The use of intersecting intervals offers little improvement over k=1. A check of the interval case frequencies in Table 6 shows that while case 5 occurs less frequently for the larger n, it is still dominant for k=0.

Table 6

Case Frequency Comparison for Θ =.9

Case	Interval	k=0 n=8192	k=0 n=16384	k=1 n=8192	k=1 n=16384
1	$[a_1,a_2]$	1	10	0	0
2	$[a_1,1]$	0	13	0	0
3	$[0,a_2]$	1	0	100	100
4	$[0,a_1]\ [a_2,1]$	7	5	0	0
5	$[0,a_1]$	91	72	0	0
6	$[a_2,1]$	0	0	0	0
7	$[0,1]$	0	0	0	0

9. Recommendations

The results in Section 8 offer an encouraging picture for fractile estimation and provides evidence on how to judge computed interval estimates for their usefulness. Based on these results one can recommend the following steps:

1. Use the computing schemes in Sections 6 and the criterion (27) in Section 7.
2. If the intervals are case 1,2 or 3 for k=0 and k=1, use these intervals and, if desired, form the intersecting interval (with lower probability) as in Section 3.

3. If cases 4, 5, 6 or 7 arise for $k=0$ do not use the interval. Do likewise for $k=1$.

4. If the number of renewals turns out to be small do not use the intervals since the applicability of asymptotic results remains in question.

10. References

1. Billingsley, P., Statistical Inference for Markov Processes, University of Chicago Press, 1961.

2. Bliss, C. I., "The Calculation of the Dosage - Mortality Curve", Annals of Applied Biology, Vol.22, 1935, pp. 134-167.

3. Bliss, C. I., "The Comparison of Dosage Mortality Data", Annals of Applied Biology, Vol. 22, 1935, pp. 307-303.

4. Blomquist, N., "The Covariance Function of the M/G/1 Queueing System", Skandivanisk Aktuavietid skrift, Vol. 50, 1967, pp. 157-174.

5. Cox, D. R., Renewal Theory, Methuen 1962.

6. Cox, D. R. and W. L. Smith, Queues, Methuen, 1961.

7. Crane, M. and D. L. Iglehart, "Simulating Stable Stochastic Systems I: General Multiserver Queues", J. ACM , Vol.21, No.1, January 1974, pp. 103-113.

8. Crane, M. and D. L. Iglehart, "Simulating Stable Stochastic Systems II: General Multiserver Queues", J. ACM, Vol. 21, January 1974, pp. 114-123.

9. Feller, W., "Fluctuation Theory of Recurrent Events", Trans. Amer. Math. Soc., Vol. 67, 1949, pp. 98-119.

10. Fieller, E. C. , The Biological Standardization of Insulin", J. Roy Stat. Soc., Supplement to Vol. 7, 1940, pp. 1-64.

11. Fishman, G. S., Concepts and Methods in Discrete Event Simulation, Wiley, 1973.

12. Miller, R. G., Simultaneous Statistical Inference, McGraw-Hill, 1966.

13. Smith, W. L., "Regenerative Stochastic Processes", Proc. Roy. Soc., Series B, Vol. 232, 1955, pp. 6-31.

Appendix

The appendix contains algorithms for computing point and $1-\alpha$ confidence interval estimates for $\Theta = \Pr\{W_1 \leq w\}$ for specified w, α and sample size n.

RUNS converts a sample record W(1),...,W(n) to a sequence of run lengths L(1),...,L(M) , where $M \leq n$, and computes a point estimate of Θ.

INPUT

JMAX ≡ upper bound on M to be considered.
FRACTW ≡ w.
n ≡ number of observations in original W record.
{W(i)} ≡ sample record.

OUTPUT

THETA ≡ point estimate of Θ.
JMAX ≡ largest value of M used.
n ≡ actual number of observations used.
{L(j)} ≡ sequence of run lengths.

KUTS computes the cutpoints for rule m^{**} in (27).

INPUT

JMAX, n, {L(j)} from RUNS output.

$$k \equiv \begin{cases} 0 \text{ if recurrent state is all zeros,} \\ 1 \text{ if recurrent state is all ones.} \end{cases}$$

OUTPUT

JM ≡ number of cutpoints ≤ JMAX/2 + 1 .
{M(j)} ≡ sequence of cutpoints for j=1,...,JM .
{COUNT(i)} ≡ frequency distribution of run lengths for k's.

NOTE: This algorithm defines M(0) ≡ 0 and can be dropped if the zero subscript is not permitted. The COUNT sequence, although not required, may assist one in determining the degree of confidence to have in the overall interval estimation procedure. See Section 9.

CALC computes the sample variances described in Section 6.

INPUT

JMAX and $\{L(j)\}$ from RUNS output.
$m \equiv$ length of the recurrent state y^*.
$k \equiv$ from KUTS input, $k=0,1$.

OUTPUT

$C \equiv$ sum of the C_j as defined in (22).
$S \equiv$ sum of the S_j as defined in (22).
$SC \equiv$ sample variance of the C_j, $s^2(C)$ as defined in (23).
$SS \equiv$ sample variance of the S_j, $s^2(S)$ as defined in (23).
$SCS \equiv$ sample covariance of C_j and S_j, $s(C,S)$ as defined in (23).
$N \equiv N_n - 1$ as defined in (22).

NOTE: This algorithm may be used for any m between 1 and M(JM), where M(JM) may be found from the output of KUTS. For $m=0$, which is the assumption of independence, one should not use this routine but notice that: C=n, S=nTHETA, SC=SCS=0, N=n and SS=nTHETA(1-THETA), where THETA is taken from the output of RUNS.

CI calculates the endpoints of the $1 - \alpha$ confidence interval for θ and estimates the variance of Z defined in (7).

INPUT

THETA $\equiv$ point estimate for θ from the output of RUNS.
ZALPHA $\equiv$ the $(1-\alpha/2)$ quantile of the standard normal distribution.
C, S, SC, SS, SCS from the output of CALC.

OUTPUT

$SZ \equiv s^2(Z)$ as defined in (25).
$A1 \equiv$ lower limit of confidence interval.
$A2 \equiv$ upper limit of confidence interval.
$B2 \equiv$ coefficient of quadratic term of $f(\theta)$ in Table 2.
DISC $\equiv$ discriminant for $f(\theta)$ in Table 2.

NOTE: If DISC is negative then one has Case 7 of Table 2. If B2 is negative one has an inverted interval, Cases 4,5 or 6 of Table 2.

Algorithm RUNS

1. $j \leftarrow 1$.
2. $L(j) \leftarrow 0$.
3. $j \leftarrow j+1$.
4. If $j \leq$ JMAX go to 2.
5. $j \leftarrow 1$.
6. XOLD $\leftarrow 0$.
7. $L(1) \leftarrow -1$.
8. $i \leftarrow 1$.
9. THETA $\leftarrow 0$.
10. $L(j) \leftarrow L(j)+1$.
11. XNEW $\leftarrow 0$.
12. If $W(i) \leq$ FRACTW then XNEW $\leftarrow 1$.
13. THETA $\leftarrow$ THETA + XNEW.
14. $j \leftarrow j + (\text{XNEW-XOLD})^2$.
15. $i \leftarrow i+1$.
16. XOLD $\leftarrow$ XNEW.
17. If $j \leq$ JMAX and $i \leq n$ go to 10.
18. JMAX $\leftarrow j-1$.
19. $n \leftarrow i-1$.
20. THETA $\leftarrow$ THETA/n.
21. RETURN.

<u>Algorithm KUTS</u>

1. $i \leftarrow 0$.
2. $i \leftarrow i+1$.
3. $COUNT(i) \leftarrow 0$.
4. If $i < n$ go to 2.
5. $j \leftarrow -1$.
6. $j \leftarrow j+1$.
7. $COUNT(L(2j+k+1)) \leftarrow COUNT(L(2j+k+1)) + 1$.
8. If $j <$ i.p.$[(JMAX-k-1)/2]$ go to 6.
9. $M(0) \leftarrow 0$.
10. $j \leftarrow 0$.
11. $i \leftarrow 0$.
12. $i \leftarrow i+1$.
13. If $COUNT(i) = 0$ and $i < n$ go to 12.
14. If $M(j) = i$ go to 17.
15. $j \leftarrow j+1$.
16. $M(j) \leftarrow i$.
17. $j \leftarrow j+1$.
18. $M(j) \leftarrow i+1$.
19. If $i < n$ go to 12.
20. $JM \leftarrow j-1$.
21. RETURN .

Algorithm CALC

1. $j \leftarrow -1$.

2. $j \leftarrow j+1$.

3. If $L(2j+k+1) < m$ and $2j+k+1 < JMAX$ go to 2.

4. $K1 \leftarrow j$.

5. $j \leftarrow$ i.p.$[(JMAX-k-1)/2] +1$.

6. $j \leftarrow j-1$.

7. If $L(2j+k+1) < m$ and $j > K1$ go to 6.

8. $K2 \leftarrow j$.

9. $N \leftarrow -1$.

10. $S \leftarrow -m$.

11. $P \leftarrow -m$.

12. $C \leftarrow 0$.

13. $LENGTH \leftarrow -m$.

14. $R \leftarrow 0$.

15. $Q \leftarrow 0$.

16. $SS \leftarrow 0$.

17. $SC \leftarrow 0$.

18. $SR \leftarrow 0$.

19. $j \leftarrow K1 - 1$.

20. $j \leftarrow j+1$.

21. $X \leftarrow L(2j+k+1)$.

22. If $X < m$ go to 34.

23. $C \leftarrow C + X + LENGTH$.

24. $S \leftarrow S + X + P$.

25. $R \leftarrow R + Q$.

26. $N \leftarrow N + 1 + X - m$.
27. $SC \leftarrow SC + X - m + (LENGTH + m)^2$.
28. $SS \leftarrow SS + X - m + (P + m)^2$.
29. $SR \leftarrow SR + Q^2$.
30. $LENGTH \leftarrow 0$.
31. $P \leftarrow 0$.
32. $Q \leftarrow 0$.
33. $X \leftarrow 0$.
34. If $j \geq K2$ go to 41.
35. $LENGTH \leftarrow LENGTH + X$.
36. $P \leftarrow P + X$.
37. $X \leftarrow L(2j+k+2)$.
38. $LENGTH \leftarrow LENGTH + X$.
39. $Q \leftarrow Q + X$.
40. Go to 20.
41. If $N \leq 0$ go to 49.
42. $SC \leftarrow (SC - C^2/N)$.
43. $SS \leftarrow (SS - S^2/N)$.
44. $SR \leftarrow (SR - R^2/N)$.
45. $SS \leftarrow kSS + (1-k)SR$.
46. $SR \leftarrow kSR + (1-k)SS$.
47. $S \leftarrow kS + (1-k)R$.
48. $SCS \leftarrow (SC + SS - SR)/2$.
49. RETURN .

Algorithm CI

1. A1 ← 0 .
2. A2 ← 1 .
3. SZ ← SS - 2THETA(SCS) + THETA^{2}SC.
4. B2 ← C^2 - ZALPHA^{2}SC .
5. B1 ← 2(C S - ZALPHA^{2}SCS).
6. B0 ← S^2 - ZALPHA^{2}SS .
7. DISC ← B1^2 - 4B0 B2 .
8. If B2 = 0 go to 19.
9. If DISC < 0 go to 18.
10. A1 ← (B1 - DISC$^{\frac{1}{2}}$)/(2B2).
11. A2 ← (B1 + DISC$^{\frac{1}{2}}$)/(2B2) .
12. If B2 > 0 go to 16.
13. TEMP ← A1.
14. A1 ← A2.
15. A2 ← TEMP.
16. If A1 < 0 then A1 ← 0 .
17. If A2 > 1 then A2 ← 1 .
18. RETURN .
19. If B1 > 0 then A1 ← B0/B1.
20. If B1 < 0 then A2 ← B0/B1.
21. Go to 16.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 14 TR-77-2

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): 6 Estimating the Mean of a Correlated Binary Sequence with an Application to Discrete Event Simulation.

5. TYPE OF REPORT & PERIOD COVERED: 9 Technical Rep[redacted]t.

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): 10 George S. Fishman [redacted] Louis R. Moore

8. CONTRACT OR GRANT NUMBER(s): 15 N00014-76-C-0302

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
University of North Carolina
Chapel Hill, N.C. 27514

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS:
Operations Research Program
Office of Naval Research
Arlington, Virginia

12. REPORT DATE: 11 Apr[redacted] 1977

13. NUMBER OF PAGES: 27 12 32p.

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of this document is unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Binary Sequence	Recurrent Events
Fractiles	Runs
Interval Estimation	Simulation
Markov chains	

Theta

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper discusses a procedure for interval estimation of the mean θ of a correlated binary (0,1) sequence. The method assumes that the sequence is strictly stationary and that a particular string of m binary digits is a recurrent event in the sequence, where $m>1$ is unknown. Of the 2^m choices for the possible recurrent events, the strings of all zeros and of all ones are examined.

m > or = to 1 2 m squared

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED 259 500 B

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

For each m=1,2,... the sequence is demarcated by entrance to the recurrent event. The subsequences between the demarcation points thus form independent epochs by assumption. Classical techniques then yield variance estimates for the number of ones and zeros in the epoch as well as an estimate of the covariance of the ones and zeros. A quadratic equation in Θ is solved to obtain an interval estimate.

Each string of all ones or all zeros examined yields a $1-\alpha$ confidence interval. The intervals are intersected to obtain shorter intervals with confidence greater than $1-2\alpha$. Since each m=1,2,... yields an interval, a conservative rule is developed to determine the m whose interval is finally used. This rule is based upon the empirical run lengths in the binary sequence.

This procedure is then applied to interval estimation of the fractile for the waiting time distribution in a simulation of the M/M/1 queue with activity level 0.9. For Θ=0.1 and 0.5 the proposed method worked well. For Θ=0.9 results showed some degradation. An error analysis led to a set of recommendations for keeping performance in practice close to the desired theoretical levels. An appendix describes algorithms for computing the critical quantities upon which the proposed method relies.